Amendments to the Specification

Please replace the paragraph at page 1, line 12 regarding Government Funding with the following amended paragraph:

Government Funding

The present invention was made with government support under Grant No. DE-FG03-00ER63041 and DE-FG02-04ER63826, awarded by the Department of Energy. The Government has certain rights in this invention.

Please replace the paragraph beginning at page 2, line 11, with the following amended paragraph.

Analytical tools that use statistically derived matrices based on allowed substitution of amino acids, are not designed to detect conservation of physical-chemical properties. For example, such tools include those available under the trademark or the trade designation of FASTA, PSI-BLAST, or BLOCKS, such as described in Pearson W., Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods in Enzymology*, 1990, 183:63-98; Schaffer et al., Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res*, 2001, 29(14):2994-3005; Schaffer et al., IMPALA: matching a protein sequence against a collection of PSI-BLAST constructed position-specific score matrices, *Bioinformatics*, 1999, 15:1000-1011; Altschul et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 1997, 25(17):3389-3402; and Henikoff et al., Increased converage of protein families with the blocks database servers, *Nucleic Acids Res*, 2000, 28:228-230.

Please replace the paragraph beginning at page 16, line 16, with the following amended paragraph.

The superiority of the novel method according to the present invention when compared to current methods for determining sequence similarity is evidenced by identifying distant homologues of the human DNA repair enzyme, apurinic/apyrimidinic endonuclease 1 (APE1) (see the Example provided herein). All the commonly used methods for genome sequence searching rely on similar, statistically derived, scoring matrices. Frequently, the same scoring matrix is used to search for related sequences (for example, with a tool available under the trademark of BLAST), to prepare a multiple alignment of the protein sequences, to analyze sequence conservation, and finally, to locate distant relatives of the family according to motif conservation. The present invention, using PCP motifs, provides an alternative and an independent way to identify distantly related proteins based on sequence information.

Please replace the paragraph beginning at page 16, line 26, with the following amended paragraph.

For example, when the human APE1 sequence was used as a query for a search using a tool available under the trade designation of PSI-BLAST ~~search~~ in the ASTRAL40 structural database with default parameters, neither known homologue of this enzyme in the database, bovine DNase-I or synaptojanin, a member of the Inositol 5'-polyphosphate phosphatase (IPP) family, was revealed. A PCP motif search according to the present invention showed, as the highest scoring proteins, all members of the DNase-I like SCOP-superfamily of APE in that database demonstrating that the present invention can find non-trivial relationships between distantly related members within superfamilies. Other high scoring proteins were from different SCOP classifications but shared functions with the APE/DNase-I/IPP superfamily, including phosphatase activity and/or metal ion binding.

Please replace the paragraph beginning at page 18, line 26, with the following amended paragraph.

One or more various embodiments of the illustrative sequence data analysis method 30, shown generally in Figure 2, shall be described with reference to Figures 3-8. The provision of the multiple sequence alignment (block 32), shown generally in Figure 2, may be provided by any suitable alignment tool. For example, such alignment tools include those available under the trademark of BLAST available on the internet through the National Center for Biotechnology Information (NCBI) (website .ncbi.nlm.nih.gov/BLAST), those available under the trade designation of CLUSTALW available on the internet through the European Bioinformatics Institute (EBI) (website ebi.ac.uk/clustalw/), or any other suitable multiple alignment tool.

Please replace the paragraph beginning at page 19, line 3, with the following amended paragraph.

One embodiment of providing a multiple sequence alignment 32 is shown in Figure 3. For example, diversely related sequences may be collected (block 50) by any suitable tool such as that available under the trade designation of BLASTP available on the internet through the National Center for Biotechnology Information (NCBI), or any other suitable tool. In other words, for example, related sequences of a known protein family desired to be used for query of a sequence database may be collected. Thereafter, a multiple sequence alignment is generated (block 52) for the related sequences of the family such as with an alignment tool as described herein. Further, the multiple sequence alignment is provided for use by the PCP motif generation program (block 54).

Please replace the paragraph beginning at page 32, line 31, with the following amended paragraph.

PCP motifs were generated for the APE protein family. Homologues of human APE1 with E-values less than 0.001 were identified in the NCBI protein sequence database using a search engine available under the trade designation of ~~the~~ BLASTP ~~search engine~~ (Altschul, et al., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402). Sequences from 42 organisms ranging from prokaryotes to eukaryotes were selected (see Figure 14) after discarding hypothetical APE-like proteins. The taxonomic classification was used to avoid excessive redundancy. Sequences were aligned with a tool available under the trade designation of CLUSTALW release 1.8 (Higgins, et al., 2000, Multiple sequence alignment, *Methods Mol. Biol.*, **143**, 1-18) using the GONNET similarity matrix (Benner, et al., 1994, Amino acid substitution during functionally constrained divergent evolution of protein sequences, *Protein Eng.*, **7**, 1323-1332), with an opening gap penalty of 10.0 and gap extension penalty of 0.2. The sequence alignment was used as input for the PCP generation program for motif identification, as shown and described with reference to Figure 5.

Please replace the paragraph beginning at page 33, line 26, with the following amended paragraph.

The sensitivity of the present invention to find proteins related to the APEs in the ASTRAL40 database were compared with that of two versions of PSI-BLAST (a trade designation for the tool used) with default parameters (E-value of 0.005), one locally installed (v 2.2.1) and the other on the web at NCBI (Altschul, et al., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402; Schaffer, et al. , 2001, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res.*, **29**, 2994-3005). To enhance the ability of a tool available under the trade designation of PSIBLAST to build a profile, the 42 sequences from the APE family,

used in the construction of the motifs, were added to the 3635 sequences from the ASTRAL40 database. The human APE sequence was used as query and ran up to five iterations using a tool available under the trade designation of PSI-BLAST. A search for APE related sequences in the BLOCKS database with the default search engine (Henikoff, et al., 2000, Increased coverage of protein families with the Blocks Database servers, *Nucleic Acids Res.*, **28**, 228-230; Henikoff, et al., 1999, Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations, *Bioinformatics*, **15**, 471-479; Henikoff, et al., 1994, Protein family classification based on searching a database of blocks, *Genomics*, **19**, 97-107) was also performed.

Please replace the paragraph beginning at page 37, line 4, with the following amended paragraph.

Identification of members of a superfamily using the currently available sequence profile methods is difficult. For example, PSI-BLAST (a trade designation for a tool used) searching, using a local program or NCBI web-based version with default parameters (E-value 0.005), detected members of the APE family in the non-redundant sequence database. However, neither version revealed DNase-I or IPP sequences even after several iterations. When the E-value was increased to 0.1, synaptojanin was revealed within the first iterations, but bovine DNase-I was only detected after 4 iterations, along with more than 500 additional entries. PSI-BLAST (a trade designation for a tool used) also failed to recognize DNase-I or synaptojanin in the ASTRAL40 database, even when we added APE sequences to allow it to form a profile. The BLOCKS (a trade designation for a tool used) search engine did not recognize homology to DNase-I even when the E-value cutoff was extended to 100. In contrast, our method identified these proteins clearly in the structural database (see Figures 9B and 9C).

Please replace the paragraph beginning at page 37, line 27, with the following amended paragraph.

All the commonly used methods for genome sequence searching rely on similar, statistically derived scoring matrices (Altschul, et al., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402; Kostich, et al., 2002, Human members of the eukaryotic protein kinase family, *Genome Biology*, **3**, 43). Frequently, the same scoring matrix is used to search for related sequences (for example, with a tool available under the trademark of BLAST), prepare a multiple alignment to analyze sequence conservation and to locate distant relatives of the family according to motif conservation. According to, Venkatarajan, et al., 2001, New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties, *J. Mol. Model.*, **7**, 445-453, the five property vectors represent all known physical-chemical properties, and provide an alternative to using the amino acid alphabet (Rigoutsos, et al., 2002, Dictionary-driven protein annotation, *Nucleic Acids Res.*, **30**, 3901-3916) or selected physical-chemical properties (Dubchak, et al., 1999, Recognition of a protein fold in the context of the SCOP classification, *Proteins*, **35**, 401-407) to identify homology. Our PCP motifs complement existing methods for functional cross networking of protein families (Marcotte, et al., 1999, A combined algorithm for genome-wide prediction of protein function, *Nature*, **402**, 83-86; Marcotte, E.M., 2000, Computational genetics: finding protein function by nonhomology methods, *Curr. Opin. Struct. Biol.*, **10**, 359-365; Overbeek, et al., 1999, The use of gene clusters to infer functional coupling, *Proc. Natl Acad. Sci. USA*, **96**, 2896-2901).